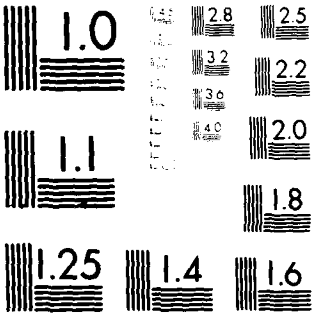
AD-A115 313 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
THE EXACT DISTRIBUTIONS OF SOME KOLMOGOROV-SMIRNOV TYPE K-SAMPL--ETC(U)
MAY 82 S G MOHANTY, B R HANDA N00014-76-C-0475
UNCLASSIFIED TR-318 NL

1 OF 1
AD A
115313

END
DATE
FILMED
07-82
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

The Exact Distributions of Some Kolmogorov-Smirnov Type k-sample Statistics

By

S. G. Mohanty and B.R. Handa

TECHNICAL REPORT NO. 318

May 6, 1982

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

1. Introduction

Let $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, k$ be k independent random samples from populations with distribution functions $F_1, \ldots, F_k$ respectively. Suppose the null hypothesis is $H_0 : F_1 = \ldots = F_k$, the hypothesis of homogeneity. For various alternatives, test statistics of Kolmogorov-Smirnov (briefly, K-S) type have been suggested by several authors (see Birnbaum and Hall (1960), Conover (1965), (1967), Dwass (1960), Kiefer (1959), and Wallenstein (1980)). In this note we consider some K-S type statistics. For any set of constants $\{c_{ij}\}$ $i = 1, \ldots, m$ (m being a positive integer), $j = 1, \ldots, k$, with $c_{i1} \neq 0$, we define two classes of K-S type statistics as follows:

(a) Two-sided statistics:

$$U(k) = \max_i \left\{ \sup_x \left| \sum_{j=1}^{k} c_{ij} F_j^*(x) \right| \right\} ,$$

(b) One-sided statistics:

$$U^+(k) = \max_i \left\{ \sup_x \left(\sum_{j=1}^{k} c_{ij} F_j^*(x) \right) \right\} ,$$

where F_j^* is the empirical distribution function of the jth sample. As an example of the use of these statistics, consider the alternatives that one of the distribution functions, say F_1, is different from the rest or greater (less) than the rest, i.e.,

$$H_1 : F_1 \neq F_i , \quad i = 2, \ldots, k ,$$

$$H_2 : F_1 > (<) F_i , \quad i = 2, \ldots, k .$$

Then one may use K-S type statistics $U_1(k) = \max_i\{\sup_x |F_1^*(x) - F_i^*(x)|\}$ and $U_1^+(k) = \max_i\{\sup_x(F_1^*(x) - F_i^*(x))\}(U_1^-(k) = \max_i\{\sup_x(F_i^*(x) - F_1^*(x))\})$

for testing against H_1 and H_2 respectively. Observe that the first one is a special case of (a) whereas the second is a special case of (b). In both cases $c_{i1} \neq 0$.

When $m = 1$, we get $\sup\limits_x \left| \sum\limits_{j=1}^{k} c_{ij} \, F_j^*(x) \right|$ from (a) and $\sup\limits_x \left(\sum\limits_{j=1}^{k} c_{ij} \, F_j^*(x) \right)$ from (b). Once we ascertain these distributions which are discrete in nature, then distributions of $\sum\limits_i \sup\limits_x \left| \sum\limits_{j=1}^{k} c_{ij} \, F_j^*(x) \right|$ and $\sum\limits_i \sup\limits_x \left(\sum\limits_{j=1}^{k} c_{ij} \, F_j^*(x) \right)$ for finite sums can be determined in a routine way. Thus distributions of statistics $U_2(k) = \sum\limits_i \sup\limits_x |F_1^*(x) - F_i^*(x)|$ for testing against H_1 and $U_2^+(k) = \sum\limits_i \sup\limits_x (F_1^*(x) - F_i^*(x))(U_2^-(k) = \sum\limits_i \sup(F_i^*(x) - F_1^*(x)))$ for testing against H_2 can be derived from a set of two-sample statistics $U_1(2)$, $U_1^+(2)$ $(U_1^-(2))$ respectively.

Let us consider testing the goodness-of-fit hypothesis that the samples are from a specified distribution with distribution function F_0. It is well known that if we apply the transform $X^* = F_0(X)$ to the original data, the null hypothesis becomes $H_0^* : F_1(x) = \ldots = F_k(x) = x$ for $x \in (0,1)$, i.e., every distribution is uniform over $(0,1)$. Analogous to (a) and (b), define the following K-S type statistics for $c_{i,k+1} \neq 0$:

(a*) Two-sided statistics:

$$V(k) = \max_i \left\{ \sup_x \left| \sum_{j=1}^{k} c_{ij} \, F_j^*(x) + c_{i,k+1} \, x \right| \right\},$$

(b*) One-sided statistics:

$$V^+(k) = \max_i \left\{ \sup_x \left(\sum_{j=1}^{k} c_{ij} \, F_j^*(x) + c_{i,k+1} \, x \right) \right\}.$$

In particular, statistics $V_1(k) = \max\limits_i \{ \sup\limits_x | x - F_i^*(x) | \}$ and $V_1^+(k) = \max\limits_i \{ \sup\limits_x (x - F_i^*(x)) \}$ may be used when the alternatives are $H_1^* : x \neq F_i(x)$, $i = 1, \ldots, k$ and $H_2^* : x > F_i(x)$, $i = 1, \ldots, k$ respectively. Notice that we have set $c_{i,k+1} = 1$. Discussion similar to the last paragraph can also be dealt with.

In this note, we derive the distributions of $U(k)$, $U^+(k)$, $V(k)$, $V^+(k)$ and the special cases.

2. Auxiliary Results

For completion, we present below the combinatorial result on multidimensional lattice paths in Handa and Mohanty (1979), for which some definitions and notations are needed. Consider (k+1)-dimensional lattice paths from the origin to the point $(n_o,n_1,\ldots,n_k)$. By the $\underline{r}(=(r_1,\ldots,r_k))$th level we mean the set of points $\{(x_o,n_1-r_1,\ldots,n_k-r_k : 0 \leq x_o \leq n_o)\}$. Denote by $a(\underline{r})$ and $b(\underline{r})$ the upper and lower restrictions at the $\underline{r}$th level, by which we mean the path at the $\underline{r}$th level can pass only through points in the set

$$\{(x_o,n_1-r_1,\ldots,n_k-r_k) : 0 \leq b(\underline{r}) \leq n_o-x_o \leq a(\underline{r}) \leq n_o\} .$$

The sets

$$A(\underline{n}) = \{a(\underline{r}) : \underline{0} \leq \underline{r} \leq \underline{n}\}$$

and

$$B(\underline{n}) = \{b(\underline{r}) : \underline{0} \leq \underline{r} \leq \underline{n}\}$$

are respectively called the upper and lower restrictions on the path. The order relation $\underline{x} \leq \underline{y}$ means $x_i \leq y_i$ for each i. Note that $a(\underline{r})$ and $b(\underline{r})$ are non-negative integers and nondecreasing in each coordinate. Also if $a(\underline{r}) < b(\underline{r})$, the path cannot pass through that level.

Let $\underline{x} \,\alpha\, \underline{y}$ mean $x_i < y_i$ for at least one i. For example, the lexicographic ordering $(\underline{u}_1,\ldots,\underline{u}_d)$ of the set $\{\underline{r} : \underline{0} \leq \underline{r} \leq \underline{n}\}$ such that $d = \prod_{i=1}^{k} (n_i+1)$, $\underline{u}_1 = 0$ and $\underline{u}_d = \underline{n}$ is an α ordering in the sense that $\underline{u}_1 \alpha \ldots \alpha \underline{u}_n$. Remember that the sequence ((0,0), (0,1), (1,0), (1,1), (2,0), (2,1)) is a lexicographic ordering of vectors $\{r : (0,0) \le (r_1, r_2) \le (2,1)\}$.

Denote by $g_k(A(\underline{n})|B(\underline{n}))$ the number of paths with upper restriction $A(\underline{n})$ and lower restriction $B(\underline{n})$. Let $\underline{n}! = \prod_{i=1}^{k} n_i!$,

$$\binom{a}{\underline{n}} = \binom{a}{n_1, \ldots, n_k} = \frac{a(a-1)\ldots\left(a - \sum_{i=1}^{k} n_i + 1\right)}{\underline{n}!}$$

$$\binom{a}{z}_+ = \binom{\max(a,o)}{z} \quad \text{and} \quad \langle \underline{n} \rangle = \sum_{i=1}^{k} n_i \, .$$

Proposition 1.

$g_k(A(\underline{n})|B(\underline{n}))$ satisfies the recurrence relation

$$\sum_{\underline{0} \le \underline{r} \le \underline{n}} (-1)^{\langle \underline{n}-\underline{r} \rangle} \binom{a(\underline{r}) - b(\underline{n}) + 1}{\underline{n} - \underline{r}}_+ g_k(A(\underline{r})|B(\underline{r})) = \delta_{\underline{o}}^{\underline{n}} \tag{1}$$

where

$$\delta_{\underline{o}}^{\underline{n}} = 1 \quad \text{when } \underline{n} = \underline{0} \, ,$$

$$= 0 \quad \text{otherwise.}$$

An explicit solution of (1) is the following:

$$g_k(A(\underline{n})|B(\underline{n})) = (-1)^{d-\langle \underline{n} \rangle - 1} \left\| \binom{a(\underline{u}_i) - b(\underline{u}_{j+1}) + 1}{\underline{u}_{j+1} - \underline{u}_i}_+ \right\|_{(d-1) \times (d-1)} \tag{2}$$

where $d = \prod_{i=1}^{k} (n_i+1)$, $\|c_{ij}\|_{m \times m}$ is the $m \times m$ determinant with (i,j)th element c_{ij} and $\{\underline{u}_1, \ldots, \underline{u}_d\} = \{\underline{r} : \underline{0} \le \underline{r} \le \underline{n}\}$ such that $\underline{0} = \underline{u}_1 \alpha \ldots \alpha \underline{u}_d = \underline{n}$.

The explicit expression (2) may be obtained by first using Cramer's rule to the system (1) of linear equations and then simplifying it. For practical purposes, the lexicographic ordering of vectors is good enough. It is easily seen that if we take $(n_o,n_1,\dots,n_k)$ as the origin and reverse the steps in the path, the upper and lower restrictions repectively become

$$A'(\underline{n}) = \{n_o - b(\underline{n} - \underline{r}) : \underline{0} \leq \underline{r} \leq \underline{n}\}$$

and

$$B'(\underline{n}) = \{n_o - a(\underline{n} - \underline{r}) : \underline{0} \leq \underline{r} \leq n)\} \;.$$

In other words, the upper and lower restrictions $a'(\underline{r})$ and $b'(\underline{r})$ are such that any path can only pass through points $\{(x_o,r_1,\dots,r_k) : 0 \leq b'(\underline{r}) \leq x_o \leq a'(\underline{r}) \leq n_o\}$ for every $\underline{r}$ in the original coordinate system. Then we get an alternative expression for the same number of paths as $g_k(A'(\underline{n})|B'(\underline{n}))$. It may be noted that for computational purposes, recurrence relation (1) is more often useful than explicit expression (2) although the determinant contains a lot of zeros.

Next we formulate a continuous analogue of Proposition 1 which gives a generalization of Steck's result (1971). Though Steck has stated the result in terms of order statistics, its connection with paths is explained in Mohanty (1979) chapters 2 and 4. In (k+1)-dimension with axes x_o, x_1, ..., x_k consider paths (not necessarily lattice paths) from the origin to $(n_o,n_1,\dots,n_k)$ where n_o is a non-negative real number and n_1, ..., n_k are non-negative integers. In this case, a path is like a lattice path except that the number of units moved at any time on x_o-axis is a non-negative real number. As before, we may define the level $\underline{r}$, the upper restriction $A(\underline{n})$ and the lower restriction $B(\underline{n})$. Here, we may remember that $a(\underline{r})$ and $b(\underline{r})$ are non-negative real numbers. In the next assertion, we adopt some of the earlier notations.

Proposition 2.

Let $g_k^*(A(\underline{n})|B(\underline{n}))$ be the measure of the set of paths with upper restriction $A(\underline{n})$ and lower restriction $B(\underline{n})$. Then $g_k^*(A(\underline{n})|B(\underline{n}))$ satisfies the recurrence relation

$$\sum_{\underline{0 < r < n}} (-1)^{\langle \underline{n}-\underline{r}\rangle} \frac{(a(\underline{r}) - b(\underline{n}))_+^{\langle \underline{n}-\underline{r}\rangle}}{(\underline{n} - \underline{r})!} g_k^*(A(\underline{r})|B(\underline{r})) = \delta_{\underline{o}}^{\underline{n}} \tag{3}$$

where $(x)_+ = \max(0,x)$. An explicit solution of (3) is given by

$$g_k^*(A(\underline{n})|B(\underline{n}))$$

$$= (-1)^{d-\langle \underline{n}\rangle-1} \left\| \frac{(a(\underline{u}_i) - b(\underline{u}_{j+1}))_+^{\langle \underline{u}_{j+1} - \underline{u}_i\rangle}}{(\underline{u}_{j+1} - \underline{u}_i)!} \right\|_{(d-1)\times(d-1)} . \tag{4}$$

The proof is inductive, follows the similar line as in Proposition 1, and therefore is omitted. Remarks following Proposition 1 are also true here. In our application we will use the alternative expressions in both cases. Without ambiguity, instead of $A'(\underline{n})$ and $B'(\underline{n})$ we may use $A(\underline{n})$ and $B(\underline{n})$ with corresponding $a(\underline{r})$ and $b(\underline{r})$.

3. Distributions

As in the two-sample case, we pool all samples and order the observations from the smallest to the largest. Let $Z_j = i-1$ if the jth member ($j = 1,\ldots,n_1+\ldots+n_k$) in the ordered sequence is $X_{i\ell}$, $\ell = 1, \ldots, n_i$, $i = 1, \ldots, k$. It can be proved that

$$P(Z_1 = z_1, \ldots, Z_{n_1+\ldots+n_k} = z_{n_1+\ldots+n_k}) = \binom{n_1+\ldots+n_k}{n_1, \ldots, n_k}^{-1} \qquad (5)$$

in the same way as in the case of $k = 2$. Represent the sequence of Z's as a k-dimensional lattice path from the origin to $(n_1, \ldots, n_k)$ by putting a unit on x_i axis whenever $Z_j = i$, $i = 0, \ldots, k-1$, $j = 1, \ldots, n_1+\ldots+n_k$. Because of (5) each path is equally likely. Also, by the virtue of their definitions statistics $U(k)$ and $U^+(k)$ can be conveniently represented as characteristics of these lattice paths. It can be verified without much difficulty that for $c > 0$,

$$P(U(k) \leq c) = M\binom{n_1+\ldots+n_k}{n_1, \ldots, n_k}^{-1}$$

and for any c,

$$P(U^+(k) \leq c) = N\binom{n_1+\ldots+n_k}{n_1, \ldots, n_k}^{-1}$$

where M is the number of paths such that a point $(x_o, x_1, \ldots, x_{k-1})$ on any path satisfies the inequalities $\left|\sum_{i=1}^{k} c_{ij}\, x_{j-1}/n_j\right| \leq c$, $i = 1, \ldots, m$ and N is the number of paths satisfying the inequalities $\left(\sum_{i=1}^{k} c_{ij}\, x_{j-1}/n_j\right) \leq c$, $i = 1, \ldots, m$. Suppose $c_{i1} > 0$. Then, using Proposition 1 (its alternative form) it turns out that $M = g_{k-1}(A(\underline{n})|B(\underline{n}))$ with

$$a(\underline{r}) = \min\left(n_1, \left\lceil \frac{n_1}{c_{i1}}\left(c - \sum_{j=2}^{k} c_{ij} \frac{r_{j-1}}{n_j}\right)\right\rceil, \ i = 1, \ldots, m\right)$$

$$b(\underline{r}) = \max\left(0, \left\lfloor \frac{n_1}{c_{i1}}\left(-c - \sum_{j=2}^{k} c_{ij} \frac{r_{j-1}}{n_j}\right)\right\rfloor, \ i = 1, \ldots, m\right)$$

where $\lceil a \rceil$ is the largest integer $\leq a$, and $\lfloor a \rfloor$ is the smallest integer $\geq a$, and $N = g_{k-1}(A(\underline{n})|B(\underline{n}))$ with $a(\underline{r})$ as above and $b(\underline{r}) = 0$. Also $\{c_{ij}\}$ must be such that the non-decreasing property of $a(\underline{r})$ and $b(\underline{r})$ has to be satisfied. This is also true in the derivation of the distributions of $V(k)$ and $V^{+}(k)$. For the example of $U_1(k)$, $a(\underline{r})$ and $b(\underline{r})$ simplify to

$$a(\underline{r}) = \min\left(n_1, \left\lceil \frac{n_1}{n_j}(r_{j-1} + cn_j) \right\rceil, j = 2,\ldots,k\right)$$

$$b(\underline{r}) = \max\left(0, \left\lfloor \frac{n_1}{n_j}(r_{j-1} - cn_j) \right\rfloor, j = 2,\ldots,k\right) .$$

When $c_{i1} < 0$, the modification is obvious and therefore is omitted.

Next let us consider the distributions of $V(k)$ and $V^{+}(k)$. Assume $c_{i,k+1} > 0$. Under the null hypothesis it can be shown by using Proposition 2 that

$$P(V(k) \leq c) = \underline{n}!\ g_k^{*}(A(\underline{n})|B(\underline{n}))$$

where

$$a(\underline{r}) = \min\left(1, \frac{1}{c_{i,k+1}}\left(c - \sum_{\substack{j=1 \\ j\neq \ell}}^{k} c_{ij}\frac{r_j}{n_j} - c_{i\ell}\frac{r_\ell + 1}{n_\ell}\right), \ell = 1,\ldots,k;\ i = 1,\ldots,m\right)$$

$$b(\underline{r}) = \max\left(0, \frac{1}{c_{i,k+1}}\left(-c - \sum_{j=1}^{k} c_{ij}\frac{r_j}{n_j}\right), i = 1,\ldots,m\right) .$$

Notice the change in $a(\underline{r})$ which is consistent with the well known expression when $k = 1$. The underlying argument is an extended version of the one-sample case. For $V^{+}(k)$, the expression is the same except that $b(\underline{r}) = 0$. In the particular case of $V_1(k)$, $a(\underline{r})$ and $b(\underline{r})$ become

$$a(\underline{r}) = \min\left(1, \frac{r_j+1}{n_j} + c, \ j = 1,\ldots,k\right)$$

$$b(\underline{r}) = \max\left(0, \frac{r_j}{n_j} - c, \ j = 1,\ldots,k\right) .$$

The expression for $V_1^+(k)$ is obtained from that of $V^+(k)$.

Professor S.G. Mohanty
McMaster University
Hamilton, Canada

Professor B.R. Handa
I.I.T.
New Delhi, India

References

Birnbaum, Z. W. and Hall, R. A. (1960). Small sample distributions for multisample statistics of the Smirnov type. *Ann. Math. Statist.* **31**, 710-720.

Conover, W. J. (1965). Several k-sample Kolmogorov-Smirnov tests. *Ann. Math. Statist.* **36**, 1019-1026.

Conover, W. J. (1967). A k-sample extension of the one-sided two-sample Smirnov test statistic. *Ann. Math. Statist.* **38**, 1726-1730.

Dwass, M. (1960). Some k-sample rank order tests. *Contributions to Probability and Statistics - Essays in Honor of Harold Hotelling.* Stanford University Press.

Handa, B. R. and Mohanty, S. G. (1979). Enumeration of higher-dimensional paths under restrictions. *Discrete Math.* **26**, 119-128.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramer-von Mises tests. *Ann. Math. Statist.* **30**, 420-447.

Mohanty, S. G. (1979). *Lattice Path Counting and Applications.* Academic Press, New York.

Steck, G. P. (1971). Rectangular probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *Ann. Math. Statist.* **42**, 1-11.

Wallenstein, S. (1980). Distribution of some one-sided k-sample Smirnov-type statistics. *J. Amer. Statist. Assoc.* **75**, 441-446.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 318	2. GOVT ACCESSION NO. AD-A115 313	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Exact Distributions of Some Kolmogorov-Smirnov Type k-sample Statistics		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) S. G. Mohanty and B. R. Handa		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office Of Naval Research Statistics & Probability Program Code 411SP Arlington, VA 22217		12. REPORT DATE May 6, 1982
		13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) k-sample tests; Kolmogorov-Smirnov type statistics; null distribution.		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) PLEASE SEE REVERSE SIDE.		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

#318

The Exact Distributions of Some Kolmogorov-Smirnov Type k-sample Statistics

Let $X_{i1}, \ldots, x_{in_i}$, $i = 1, \ldots, k$ be k independent random samples from populations with distribution functions $F_1, \ldots, F_k$ respectively. For the hypothesis of homogeneity $H_0 : F_1 = \ldots = F_k$, we obtain the distributions of two Kolmogorov-Smirnov type statistics $\max_i \left\{ \sup_x \left| \sum_{j=1}^{k} c_{ij} \; F_j^*(x) \right| \right\}$ and $\max_i \left\{ \sup_x \left(\sum_{j=1}^{k} c_{ij} \; F_j^*(x) \right) \right\}$ where $\{c_{ij}\}$ is a set of constants with $c_{i1} \neq 0$ and F_j^* is the empirical distribution function of the jth sample. Also, for the goodness-of-fit hypothesis similar statistics are considered and their distributions derived. Some special cases are discussed.